Data IDE

Product Introduction

Product Introduction

##1. Product architecture DFD

“Data IDE” is a platform product launched by Alibaba Cloud in the field of big data. It provides one-stop big data development, data permission management, offline job scheduling, and many other features. It is dependent on the massive data computing engine MaxCompute (originally ODPS, independently developed by Alibaba Cloud) in the underlayer and provides features that are applicable to multiple scenarios, including offline processing, analysis, cloud data warehouse building, and big data mining. It is offered in an ‘out-of-the-box’ manner and you don’t need to worry much about the cost and complexity of the underlying cluster establishment and O&M.

The product architecture is shown in the figure below:

From the figure above, we can see that the underlayer of Data IDE Kit is an integrated development environment based on the MaxCompute (originally ODPS). Data IDE offers metadata synchronization of massive heterogeneous data, offline scheduling, workflow configuration, MR, and machine learning capabilities. In addition, it can be seamlessly integrated with the BI, DataV and recommendations of the Alibaba Cloud to provide you with a more convenient one-stop platform.

2. Major features of Data IDE Kit

Data IDE Kit introduces a brand new workflow job design philosophy and has the following features compared with the previous version:

1) Drag-and-drop workflow interface

The system’s Data Development module provides abundant visual components, including SQL (ODPS SQL), data synchronization, MR (ODPSMR), machine learning, shell, and other job types. Compared with open-source workflow drag-and-drop operations, it provides a more convenient and flexible experience and interaction.

2) Personalized data favorites and management

The system data management module provides personalized data favorites and management. You can easily add data tables of interest to favorites, manage the lifecycle, basic information and owner of a data table, and view the storage information, partition information, output information and kinship information of the data table.

3) One-click job publishing across projects

Quick migration and publishing of jobs between different projects are provided under the same

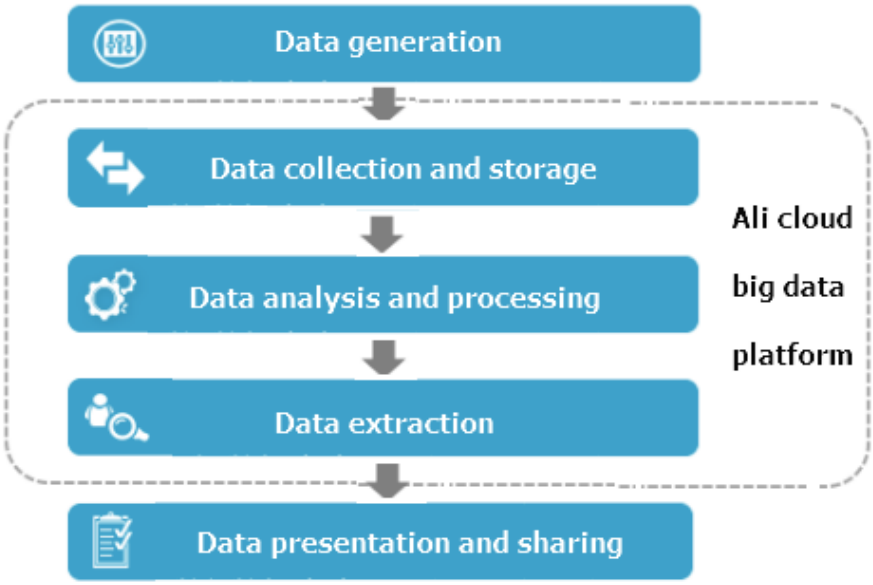
primary account. We provide a dual-environment model for customers simulating the 'development' and 'production' environments and more offline and online production models.

4) Visual job monitoring

The O&M Center provides a visual job monitoring and management tool and supports displaying the overall job running conditions in a DAG. Exception management is also more convenient. Operations such as "rerun", "restore", "suspend", and "stop" are supported.

3. Development process

The data development usually goes through the following processes under normal conditions:



The figure above shows that the overall data development process includes data generation, data collection and storage, data analysis and computing, data extraction, and data presentation and sharing. All the data development processes framed by dotted lines can be completed on the Alibaba Cloud Data IDE. Descriptions can be found below:

1) Data generation

A business system will generate a large amount of structured data every day. The data is stored in the database of the business system, including MySQL, Oracle, and RDS.

2) Data collection and storage

You need to first synchronize the data of different business systems to MaxCompute (originally ODPS) before leveraging the massive data storage and processing capabilities of MaxCompute (originally ODPS) for analyzing the existing data. The Alibaba Cloud Data IDE platform provides the data synchronization service to synchronize various types of data sources in the business system with MaxCompute (originally ODPS) according to the predefined scheduling period.

3) Data analysis and processing

Following the above step, you can start the processing (ODPS SQL and ODPS MR), analysis and mining (data analysis and data mining) of data on MaxCompute (originally ODPS) to discover the value of the data.

4) Data extraction

The result data after analysis and processing should be synchronously exported to the business systems so that the business personnel can utilize the value of the data.

5) Data presentation and sharing

The results of big data analysis and processing are presented and shared using reports and geographic information systems.

Common Data IDE scenarios under normal conditions include the following:

> Data IDE conveniently migrates data produced by the business system to the cloud, constructs large-scale data warehouses and BI applications, and uses the massive data storage and processing capabilities of ODPS.
>
> With quick data use and analysis based on Data IDE, you can export big data processing results and directly apply the results to the business system to achieve data-based operations.
>
> Data IDE provides a unified and user-friendly scheduling system and a visual O&M scheduling interface from the perspective of complex job scheduling and O&M, solving inconvenient O&M management issues.

Common actions

- New Project
- New Data Source
- Add a Project Member
- Tabulation with Script
- Visualized Tabulation
- Import Data
- New Wordflow
- New Job
- New UDF
- Use OPEN MR
- Workflow O&M

- Scheduling Parameter Usage